Summary

This analysis was conducted for X Education to identify strategies for attracting more industry professionals to enroll in their courses. The provided dataset offered extensive insights into potential customers' behavior, including their site visit patterns, time spent on the site, methods of reaching the site, and the overall conversion rate.

The following are the steps used:

1. Cleaning data:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed as below: -

- Drop columns with more than 40% missing values.
- Impute missing values for numerical columns using median.
- Impute missing values for categorical columns using mode.

2. EDA:

A quick EDA for (Univariate, Bivariate, Multivariate) Was done to check the condition of our data. After doing univariate analysis It was found that a lot of categorical variables were irrelevant. The numeric values seem good and outliers were Identify using IQR method and where treated.

3. Hypothesis Testing

The hypothesis testing results indicate that significant differences exist between converted and non-converted leads in terms of total time spent on the website, lead source, total visits, and city, suggesting these factors influence conversion. However, page views per visit and Asymmetrique Activity Score showed no significant differences, indicating they do not impact conversion rates.

4. Dummy Variables:

The dummy variables were created 'Do Not Email', 'Do Not Call' binary variables (Yes/No) was converted to 1/0. For numeric values we used the get_dummies function of pandas library.

5. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

6. Model Building:

Firstly, we did Feature Selection Using RFE and selected top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

7. Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

8. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

It was found that the variables that mattered the most in the potential buyers are (In descending order) with respect to highest coefficients (which indicate the strongest positive impact on the target variable) are:

- Lead Origin_lead add form 3.241959
- What is your current occupation_working professional 2.326409
- Total Time Spent on Website 1.048084
- Lead Source_olark chat 0.994948
- Last Activity_converted to lead -1.107568
- Specialization_hospitality management -1.115232
- Last Activity_olark chat conversation -1.173468
- Do Not Email -1.547412
- Last Notable Activity_page visited on website -1.708281
- Asymmetrique Activity Index_03.low -1.784327
- Last Notable Activity_email opened -1.791193
- Last Notable Activity_modified -1.873820
- Last Notable Activity_email link clicked -1.924711
- Last Notable Activity_olark chat conversation -1.996727
- Lead Quality_might be -2.648833
- Lead Quality_not sure -3.385105
- Lead Quality_worst -5.183895

NOTE- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

X-----X-----X------X